

On generating independent random strings

Marius Zimand¹ *

Department of Computer and Information Sciences, Towson University, Baltimore,
MD, USA

Abstract. It is shown that from two strings that are partially random and independent (in the sense of Kolmogorov complexity) it is possible to effectively construct polynomially many strings that are random and pairwise independent. If the two initial strings are random, then the above task can be performed in polynomial time. It is also possible to construct in polynomial time a random string, from two strings that have constant randomness rate.

Keywords: Kolmogorov complexity, random strings, independent strings, randomness extraction.

1 Introduction

This paper belongs to a line of research that investigates whether certain attributes of randomness can be improved effectively. We focus on finite binary strings and we regard randomness from the point of view of Kolmogorov complexity. Thus, the amount of randomness in a binary string x is given by $K(x)$, the Kolmogorov complexity of x and the randomness rate of x is defined as $K(x)/|x|$, where $|x|$ is the length of x . Roughly speaking, a string x is considered to be random if its randomness rate is approximately equal to 1. It is obvious that randomness cannot be created from nothing (e.g., from the empty string). On the other hand, it might be possible that if we already possess some randomness, we can produce “better” randomness or “new” randomness. For the case when we start with *one* string x , it is known that there exists no computable function that produces another string y with higher randomness rate (i.e., “better” randomness), and it is also clear that there is no computable function that produces “new” randomness, by which we mean a string y that has non-constant Kolmogorov complexity conditioned by x . In fact, Vereshchagin and Vyugin [VV02, Th. 4] construct a string x with high Kolmogorov complexity so that any shorter string that has small Kolmogorov complexity conditioned by x (in particular any string effectively constructed from x) has small Kolmogorov complexity unconditionally. Therefore, we need to analyze what is achievable if we start with two or more strings that have a certain amount of randomness and a certain degree of independence. In this case, in certain circumstances, positive solutions exist.

* The author is supported in part by NSF grant CCF 0634830.
<http://triton.towson.edu/~mzimand>.

For example, Fortnow, Hitchcock, Pavan, Vinodchandran and Wang [FHP⁺06] show that, for any σ there exists a constant ℓ and a polynomial-time procedure that from an input consisting of ℓ n -bit strings x_1, \dots, x_ℓ , each with Kolmogorov complexity at least σn , constructs an n -bit string with Kolmogorov complexity $\succeq n - \text{dep}(x_1, \dots, x_\ell)$ ($\text{dep}(x_1, \dots, x_\ell)$ measures the dependency of the input strings and is defined as $\sum_{i=1}^{\ell} K(x_i) - K(x_1 \dots x_\ell)$; \succeq means that the inequality holds within an error of $O(\log n)$).

In this paper we focus on the case when the input consists of *two* strings x and y of length n . We say that x and y have dependency at most $\alpha(n)$ if the complexity of each string does not decrease by more than $\alpha(n)$ when it is conditioned by the other string, i.e., if $K(x) - K(x | y) \leq \alpha(n)$ and $K(y) - K(y | x) \leq \alpha(n)$. The reader should have in mind the situation $\alpha(n) = O(\log n)$, in which case we say that x and y are independent (see [CZ08] for a discussion of independence for finite binary strings and infinite binary sequences). We address the following two questions:

Question 1. Given x and y with a certain amount of randomness and a certain degree of independence, is it possible to effectively/efficiently construct a string z that is random?

Question 2. (a more ambitious version of Question 1) Given x and y with a certain amount of randomness and a certain degree of independence, is it possible to effectively/efficiently construct strings that are random and have small dependency with x , with y , and pairwise among themselves? How many such strings exhibiting “new” randomness can be produced?

A construction is *effective* if it can be done by a computable function, and it is *efficient* if it can be done by a polynomial-time computable function.

We first recall the well-known (and easy-to-prove) fact that if x and y are random and independent, then the string z obtained by bit-wise XOR-ing the bits of x and y is random and independent with x and with y . Our first result is an extension of the above fact.

Theorem 1. (Informal statement.) If x and y are random and have dependency at most $\alpha(n)$, then by doing simple arithmetic operations in the field $\text{GF}[2^n]$ (which take polynomial time), it is possible to produce polynomially many strings $z_1, \dots, z_{\text{poly}(n)}$ of length n such that $K(z_i) \succeq n - \alpha(n)$ and the strings $x, y, z_1, \dots, z_{\text{poly}(n)}$ are pairwise at most $\approx \alpha(n)$ -dependent, where $\approx (\succeq)$ means that the equality (resp., the inequality) is within an error of $O(\log n)$. In particular, if x and y are independent, then the output strings are random and together with the input strings form a collection of pairwise independent strings.

The problem is more complicated when the two input strings x and y have randomness rate significantly smaller than 1. In this case, our questions are related to randomness extractors, which have been studied extensively in computational complexity. A randomness extractor is a polynomial-time computable procedure that improves the quality of a defective source of randomness. A source of randomness is modeled by a distribution X on $\{0, 1\}^n$, for some n , and its defectiveness is modeled by the min-entropy of X (X has min-entropy k if 2^{-k}

is the largest probability that X assigns to any string in $\{0, 1\}^n$). There are several type of extractors; for us, multi-source extractors are of particular interest. An ℓ -multisource extractor takes as input ℓ defective independent distributions on the set of n -bit strings and outputs a string whose induced distribution is statistically close to the uniform distribution. The analogy between randomness extractors and our questions is quite direct: The number of sources of the extractor corresponds to the number of input strings and the min-entropy of the sources corresponds to the Kolmogorov complexity of the input strings. For $\ell = 2$, the best multisource extractors are (a) the extractor given by Raz [Raz05] with one source having min-entropy $((1/2) + \alpha)n$ (for some small α) and the second source having min-entropy $\text{polylog}(n)$, and (b) the extractor given by Bourgain [Bou05] with both sources having min-entropy $((1/2) - \alpha)n$ (for some small α). Both these extractors are based on recent results in arithmetic combinatorics. It appears that finding polynomial-time constructions achieving the goals in Question 2 is difficult. If we settle for effective constructions, then positive solutions exist. In [Zim09], we have shown that there exists a computable function f such that if x and y have Kolmogorov complexity $s(n)$ and dependency at most $\alpha(n)$, then $f(x, y)$ outputs a string z of length $m \approx s(n)/2$ such that $K(z | x) \geq m - \alpha(n)$ and $K(z | y) \geq m - \alpha(n)$. Our second result extends the methods from [Zim09] and shows that it is possible to effectively construct polynomially many strings exhibiting “new” randomness.

Theorem 2. (Informal statement.) For every function $O(\log n) \leq s(n) \leq n$, there exists a computable function f such that if x and y have Kolmogorov complexity $s(n)$ and dependency at most $\alpha(n)$, then $f(x, y)$ outputs polynomially many strings $z_1, \dots, z_{\text{poly}(n)}$ of length $m \approx s(n)/3$ such that $K(z_i) \geq m - \alpha(n)$ and the strings $(x, y, z_1, \dots, z_{\text{poly}(n)})$ are pairwise at most $\approx \alpha(n)$ -dependent. In particular, if x and y are independent, then the output strings are random and together with the input strings form a collection of pairwise independent strings.

For Question 1, we give a polynomial-time construction in case x and y have linear Kolmogorov complexity, i.e., $K(x) \geq \delta n$ and $K(y) \geq \delta n$, for a positive constant $\delta > 0$. The proof relies heavily on a recent result of Rao [Rao08], which shows the existence of 2-source condensers. (A 2-source condenser is similar but weaker than a 2-source extractor in that the condenser’s output is only required to be statistically close to a distribution that has larger min-entropy rate than that of its inputs, while the extractor’s output is required to be statistically close to the uniform distribution.)

Theorem 3. (Informal statement.) For every constant $\delta > 0$, there exists a polynomial-time computable function f such that if x and y have Kolmogorov complexity δn and dependency at most $\alpha(n)$, then $f(x, y)$ outputs a string z of length $m = \Omega(\delta n)$ and $K(z) \geq m - (\alpha(n) + \text{poly}(\log n))$.

The main proof technique is an extension of the method used in [Zim08] and in [Zim09]. It uses ideas from Fortnow et al. [FHP⁺06], who showed that a multi-source extractor can also be used to extract Kolmogorov complexity. A key element is the use of *balanced tables*, which are combinatorial objects similar to

2-source extractors. A balanced table is an N -by- N table whose cells are colored with M colors in such a way that each sufficiently large rectangle inside the table is colored in a balanced way, in the sense that all colors appear approximately the same number of times. The exact requirements for the balancing property are tailored according to their application. The type of balanced table required in Theorem 2 is shown to exist using the probabilistic method and then constructed using exhaustive search. This is why the transformation in Theorem 2 is only effective, and not polynomial-time computable. The existence of the type of balanced table used in Theorem 3 is a direct consequence of Rao's 2-source condenser.

The paper is structured as follows. Sections 1.1 and 1.2 introduce the notation and the main concepts of Kolmogorov complexity. Section 1.3 is dedicated to balanced tables. Theorem 1 and Theorem 2 are proved in Section 2, and Theorem 3 is proved in Section 3.

1.1 Preliminaries

\mathbb{N} denotes the set of natural numbers. For $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, 2, \dots, n\}$. We work over the binary alphabet $\{0, 1\}$. A string is an element of $\{0, 1\}^*$. If x is a string, $|x|$ denotes its length. The cardinality of a finite set A is denoted $|A|$. Let M be a standard Turing machine. For any string x , define the *Kolmogorov complexity* of x with respect to M , as $K_M(x) = \min\{|p| \mid M(p) = x\}$. There is a universal Turing machine U such that for every machine M there is a constant c such that for all x , $K_U(x) \leq K_M(x) + c$. We fix such a universal machine U and dropping the subscript, we let $K(x)$ denote the Kolmogorov complexity of x with respect to U . For the concept of conditional Kolmogorov complexity, the underlying machine is a Turing machine that in addition to the read/work tape which in the initial state contains the input p , has a second tape containing initially a string y , which is called the conditioning information. Given such a machine M , we define the Kolmogorov complexity of x conditioned by y with respect to M as $K_M(x \mid y) = \min\{|p| \mid M(p, y) = x\}$. Similarly to the above, there exist universal machines of this type and a constant c and they satisfy the relation similar to the one above, but for conditional complexity. We fix such a universal machine U , and dropping the subscript U , we let $K(x \mid y)$ denote the Kolmogorov complexity of x conditioned by y with respect to U . In this paper, the constants implied in the $O(\cdot)$ notation depend only on the universal machine.

The Symmetry of Information Theorem (see [ZL70]) states that for all strings x and y :

$$|(K(x) - K(x \mid y)) - (K(y) - K(y \mid x))| \leq O(\log K(x) + \log K(y)). \quad (1)$$

In case the strings x and y have length n , it can be shown that

$$|(K(x) - K(x \mid y)) - (K(y) - K(y \mid x))| \leq 2 \log n + O(1). \quad (2)$$

Sometimes we need to concatenate two strings a and b in a self-delimiting matter, i.e., in a way that allows to retrieve each one of them. A simple way to do

this is by taking $a_1a_1a_2a_2\ldots a_na_n01b$, where $a = a_1\ldots a_n$, with each $a_i \in \{0, 1\}$. A more efficient encoding is as follows. Let $|a|$ in binary notation be $c_1c_2\ldots c_k$. Note that $k = \lfloor \log |a| \rfloor + 1$. Then we define $\text{concat}(a, b) = c_1c_1c_2c_2\ldots c_kc_k01ab$. Note that $|\text{concat}(a, b)| = |a| + |b| + 2\lfloor \log |a| \rfloor + 4$.

1.2 Independent strings

Definition 1. (a) Two strings x and y are at most $\alpha(n)$ -dependent if $K(x) - K(x|y) \leq \alpha(|x|)$ and $K(y) - K(y|x) \leq \alpha(|y|)$.

(b) The strings (x_1, x_2, \ldots) are pairwise at most $\alpha(n)$ -dependent, if for every $i \neq j$, x_i and x_j are at most $\alpha(n)$ -dependent.

1.3 Balanced tables

A table is a function $T : [N] \times [N] \rightarrow [M]$. In our applications, N and M are powers of 2, i.e., $N = 2^n$ and $M = 2^m$. We identify $[N]$ with $\{0, 1\}^n$ and $[M]$ with $\{0, 1\}^m$. Henceforth, we assume this setting.

It is convenient to view such a function as a two dimensional table with N rows and N columns where each entry has a color from the set $[M]$. If B_1, B_2 are subsets of $[N]$, the $B_1 \times B_2$ rectangle of table T is the part of T comprised of the rows in B_1 and the columns in B_2 . If $A \subseteq \{0, 1\}^m$ and $(x, y) \in [N] \times [N]$, we say that the cell (x, y) is A -colored if $T(x, y) \in A$.

In our proofs, we need the various tables to be *balanced*, which, roughly speaking, requires that in each sufficiently large rectangle $B_1 \times B_2$, all colors appear approximately the same number of times.

One variant of this concept is given in the following definition.

Definition 2. Let $k \in \mathbb{N}$. The table T is (S, n^k) -strongly balanced if for every pair of sets B_1 and B_2 , where $B_1 \subseteq [N]$, $|B_1| \geq S$, $B_2 \subseteq [N]$, $|B_2| \geq S$, the following two inequalities hold:

(1) For every $a \in [M]$,

$$|\{(x, y) \in B_1 \times B_2 \mid T(x, y) = a\}| \leq \frac{2}{M} |B_1 \times B_2|,$$

(2) for every $(a, b) \in [M]^2$ and for every $(i, j) \in [n^k]^2$,

$$|\{(x, y) \in B_1 \times B_2 \mid T(x + i, y) = a \text{ and } T(x + j, y) = b\}| \leq \frac{2}{M^2} |B_1 \times B_2|,$$

where addition is done modulo N .

Using the probabilistic method, we show that, under some settings for the parameters, strongly-balanced tables exist.

Lemma 1. If $S^2 > 3M^2 \ln M + 6M^2 \cdot k \cdot \ln n + 6SM^2 + 6SM^2 + 6SM^2 \ln(N/S) + 3M^2$, then there exists an (S, n^k) - strongly balanced table.

NOTE: The condition is satisfied if $M = o((1/\sqrt{n})S^{1/2})$.

Proof. We first fix $(a, b) \in [M]^2$, two sets B_1 and B_2 with $B_1 \subseteq [N]$, $|B_1| = S$, $B_2 \subseteq [N]$, $|B_2| = S$. Note that for a fixed cell $(x, y) \in B_1 \times B_2$ and fixed $j \in [n^k]$, $\text{Prob}[T(x, y) = a] = 1/M$ and $\text{Prob}[T(x, y) = a \text{ and } T(x + j, y) = b] = 1/M^2$.

Therefore, by the Chernoff bounds,

$$\text{Prob}\left[\frac{\text{number of } a\text{-colored cells in } B_1 \times B_2}{S^2} > 2\frac{1}{M}\right] \leq e^{-(1/3)(1/M)S^2},$$

and, for fixed j ,

$$\text{Prob}\left[\frac{\text{number of } (a, b)\text{-colored } j\text{-apart cells in } B_1 \times B_2}{S^2} > 2\frac{1}{M^2}\right] \leq e^{-(1/3)(1/M^2)S^2}.$$

There are M possibilities for choosing a , and the number of possibilities for choosing the sets B_1 and B_2 is $\binom{N}{S}^2 \leq (eN/S)^{2S} = e^{2S+2S \ln(N/S)}$. Therefore, the probability that the relation (1) in Definition 2 does not hold is bounded by

$$e^{-(1/3)(1/M)S^2 + \ln M + 2S + 2S \ln(N/S)}. \quad (3)$$

There are M^2 possibilities for choosing (a, b) , n^{2k} possibilities for (i, j) and the number of possibilities for choosing the sets B_1 and B_2 is $\binom{N}{S}^2 \leq (eN/S)^{2S} = e^{2S+2S \ln(N/S)}$. Therefore, the probability that the relation (2) in Definition 2 does not hold is bounded by

$$e^{-(1/3)(1/M^2)S^2 + 2 \ln M + 2k \ln n + 2S + 2S \ln(N/S)}. \quad (4)$$

If the parameters satisfy the requirement stated in the hypothesis, then the bound in Equation (3) is less than $e^{-1} < 1/2$ and the bound in Equation (4) is less than $e^{-1} < 1/2$. Therefore the probability that both relation (1) and relation (2) hold is positive, and thus there exists a (S, n^k) -strongly balanced table. \blacksquare

The above proof uses the probabilistic method which does not indicate an efficient way to construct such tables. In our application, we will build such tables by exhaustive search, an operation that can be done in EXPSPACE.

A weaker type of a balanced table can be constructed in polynomial-time using a recent result of Rao [Rao08]. We first recall the following definitions. Let X and Y be two probability distributions on $\{0, 1\}^n$. The distributions X and Y are ϵ -close if for every $A \subseteq \{0, 1\}^n$, $|\text{Prob}(X \in A) - \text{Prob}(Y \in A)| < \epsilon$. The min-entropy of distribution X is $\max_{a \in \{0, 1\}^n} (\log(1/\text{Prob}(X = a)))$.

Fact 1 [Rao08] *For every $\delta > 0$, $\epsilon > 0$, there exists a constant c and a polynomial-time computable function $\text{Ext} : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}^m$, where $m = \Omega(\delta n)$, such that if X and Y are two independent random variables taking values in $\{0, 1\}^n$ and following distributions over $\{0, 1\}^n$ with min-entropy at least δn , then $\text{Ext}(X, Y)$ is ϵ -close to a distribution with min-entropy $m - (\delta \log 1/\epsilon)^c$.*

Rao's result easily implies the existence of a polynomial-time table with a useful balancing property.

Lemma 2. *Let $\delta > 0$, $\epsilon > 0$ and let c be the constant and $Ext : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}^m$ be the function from Theorem 1, corresponding to these parameters. We identify $\{0, 1\}^n$ with $[N]$ and $\{0, 1\}^m$ with $[M]$ and view Ext as an $[N] \times [N]$ table colored with M colors. Then for every rectangle $B_1 \times B_2 \subseteq [N] \times [N]$, where $|B_1| \geq 2^{\delta n}$ and $|B_2| \geq 2^{\delta n}$ and for every $A \subseteq [M]$, the number of cells in $B_1 \times B_2$ that are A -colored is at most*

$$\left(\frac{|A|}{M} 2^{(\delta \log(1/\epsilon))^c} + \epsilon \right) \cdot |B_1 \times B_2|.$$

Proof. Let B_1 and B_2 be two subsets of $\{0, 1\}^n$ of size $\geq 2^{\delta n}$. Let X and Y be two independent random variables that follow the uniform distributions on B_1 , respectively B_2 and assume the value 0 on $\{0, 1\}^n - B_1$, respectively 0 on $\{0, 1\}^n - B_2$. Since X and Y have min-entropy $\geq 2^{\delta n}$, it follows that $Ext(X, Y)$ is ϵ -close to a distribution Z on $\{0, 1\}^m$ that has min-entropy $m - (\delta \log 1/\epsilon)^c$. If $A \subseteq \{0, 1\}^m$, then Z assigns to A probability mass at most $\frac{|A|}{M} 2^{(\delta \log 1/\epsilon)^c}$, because it assigns to each element in $\{0, 1\}^m$ at most $2^{-(m - (\delta \log 1/\epsilon)^c)}$. Thus, $Ext(X, Y)$ assigns to A probability mass at most $\frac{|A|}{M} 2^{(\delta \log 1/\epsilon)^c} + \epsilon$. This means that the number of occurrences of A -colored cells in the $B_1 \times B_2$ rectangle is bounded by $\left(\frac{|A|}{M} 2^{(\delta \log 1/\epsilon)^c} + \epsilon \right) \cdot |B_1 \times B_2|$. \blacksquare

2 Generating multiple random independent strings

We prove Theorem 1. The formal statement is as follows.

Theorem 1. *For every $k \in \mathbb{N}$, there is a polynomial-time computable function f that on input x_1, x_2 , two strings of length n , outputs n^k strings $x_3, x_4, \dots, x_{n^k+2}$, strings of length n , with the following property. For every sufficiently large n and for every function $\alpha(n)$, if x_1 and x_2 satisfy*

- (i) $K(x_1) \geq n - \log n$,
- (ii) $K(x_2) \geq n - \log n$, and
- (iii) x_1 and x_2 are at most $\alpha(n)$ -dependent,

then

- (a) $K(x_i) \geq n - (\alpha(n) + (k + O(1)) \log n)$, for every $i \in \{3, \dots, n^k + 2\}$, and
- (b) the strings $x_1, x_2, \dots, x_{n^k+2}$ are pairwise at most $\alpha(n) + (3k + O(1)) \log n$ -dependent.

Proof. Let $x_1, x_2 \in \{0, 1\}^n$ be such that $K(x_1) \geq n - \log n$, $K(x_2) \geq n - \log n$. Since x_1 and x_2 are at most $\alpha(n)$ -dependent, $K(x_1 \mid x_2) \geq K(x_1) - \alpha(n) \geq n - (\alpha(n) + \log n)$. Similarly, $K(x_2 \mid x_1) \geq n - (\alpha(n) + \log n)$.

The function f outputs

$$\begin{aligned} x_3 &= x_1 + 1 \cdot x_2, \\ x_4 &= x_1 + 2 \cdot x_2, \\ &\vdots \\ x_{n^k+2} &= x_1 + n^k \cdot x_2, \end{aligned}$$

where the arithmetic is done in the finite field $\text{GF}[2^n]$ and $1, 2, \dots, n^k$ denote the first (in some canonical ordering) n^k non-zero elements of $\text{GF}[2^n]$.

Let x_i be one of the “new” strings, i.e., $i \in \{3, \dots, n^k + 2\}$. Let t be defined by $K(x_i | x_1) = t$. Observe that given x_1 , i (that can be described with $k \log n$ bits) and $t + O(1)$ bits we can construct x_2 ; first we compute x_i and then from x_1 and x_i , we derive x_2 .

Therefore, $K(x_2 | x_1) \leq t + k \log n + 2(\log k + \log \log n) + O(1)$. Since $K(x_2 | x_1) \geq n - (\alpha(n) + \log n)$, it follows that $t \geq n - (\alpha(n) + (k + O(1)) \log n)$ (taking into account that $k < n$; if $k \geq n$, the theorem holds trivially). Therefore, $K(x_i | x_1) \geq n - (\alpha(n) + (k + O(1)) \log n)$ (which implies (a)). We infer that

$$\begin{aligned} K(x_i) - K(x_i | x_1) &\leq (n + O(1)) - (n - (\alpha(n) + (k + O(1)) \log n)) \\ &= \alpha(n) + (k + O(1)) \log n. \end{aligned}$$

By the Symmetry of Information Theorem, $K(x_1) - K(x_1 | x_i) \leq \alpha(n) + (k + O(1)) \log n$, and thus x_i and x_1 are at most $\alpha(n) + (k + O(1)) \log n$ -dependent.

Similarly, x_i and x_2 are at most $\alpha(n) + (k + O(1)) \log n$ -dependent. Thus, (b) follows for pairs (x_i, x_1) and (x_i, x_2) with $i \geq 3$.

Let us next consider a pair of strings (x_i, x_j) with $i \neq j$ and $i, j \in \{3, \dots, n^k + 2\}$. Let t be defined by $K(x_i | x_j) = t$. Note that given x_j , i and j and $t + O(1)$ bits we can construct x_1 : first we compute x_i and then from x_i and x_j , we derive x_1 . Therefore,

$$K(x_1 | x_j) \leq t + 2k \log n + 2(\log k + \log \log n) + O(1).$$

Recall that

$$K(x_1) - K(x_1 | x_j) \leq \alpha(n) + (k + O(1)) \log n.$$

Then,

$$\begin{aligned} t + 2k \log n + 2(\log k + \log \log n) + O(1) &\geq K(x_1 | x_j) \\ &\geq K(x_1) - (\alpha(n) + (k + O(1)) \log n) \\ &\geq n - (\alpha(n) + (k + O(1)) \log n). \end{aligned}$$

Thus, $K(x_j | x_i) = t \geq n - (\alpha(n) + (3k + O(1)) \log n)$. It follows that

$$\begin{aligned} K(x_j) - K(x_j | x_i) &\leq (n + O(1)) - (n - (\alpha(n) + (3k + O(1)) \log n)) \\ &\leq \alpha(n) + (3k + O(1)) \log n. \end{aligned}$$

Thus, x_j and x_i are at most $\alpha(n) + (3k + O(1)) \log n$ -dependent. ■

We next prove Theorem 2. The formal statement is as follows.

Theorem 2. *For every $k \in \mathbb{N}$, for every computable function $s(n)$ verifying $(6k + 15) \log n < s(n) \leq n$ for every n , there exists a computable function f that, for every n , on input two strings x_1 and x_2 of length n , outputs n^k strings $x_3, x_4, \dots, x_{n^k+2}$ of length $m = s(n)/3 - (2k+5) \log n$ with the following property. For every sufficiently large n and for every function $\alpha(n)$, if*

- (i) $K(x_1) \geq s(n)$,
- (ii) $K(x_2) \geq s(n)$ and
- (iii) x_1 and x_2 are at most $\alpha(n)$ - dependent,

then

- (a) $K(x_i) \geq m - (\alpha(n) + O(\log n))$, for every $i \in \{3, \dots, n^k + 2\}$ and
- (b) the strings in the set $\{x_1, x_2, \dots, x_{n^k+2}\}$ are pairwise at most $\alpha(n) + (2k + O(1)) \log n$ -dependent.

Proof. We fix n and let $N = 2^n$, $m = s(n)/3 - (2k + 5) \log n$, $M = 2^m$, $S = 2^{2s(n)/3}$. We also take $t = \alpha(n) + 7 \log n$. The requirements of Lemma 1 are satisfied and therefore there exists a table $T : [N] \times [N] \rightarrow [M]$ that is (S, n^k) -strongly balanced. By brute force, we find the smallest (in some canonical sense) such table T . Note that the table T can be described with $\log n + O(1)$ bits.

The function f outputs

$$\begin{aligned} x_3 &= T(x_1 + 1, x_2), \\ x_4 &= T(x_1 + 2, x_2), \\ &\vdots \\ x_{n^k+2} &= T(x_1 + n^k, x_2). \end{aligned}$$

We show the following two claims.

Claim 1 *For every $j \in \{3, \dots, n^k + 2\}$, $K(x_j \mid x_1) \geq K(x_j) - (\alpha(n) + O(\log n))$ and $K(x_j \mid x_2) \geq K(x_j) - (\alpha(n) + O(\log n))$.*

Claim 2 *For every $i, j \in \{3, \dots, n^k + 2\}$, $K(x_j \mid x_i) \geq K(x_j) - (\alpha(n) + (2k + O(1))) \log n$.*

Claim 1 is using ideas from the paper [Zim09]. For the sake of making this paper self-contained we present the proof. Let $j \in \{3, \dots, n^k + 2\}$. We show that $K(x_j \mid x_1)$ and $K(x_j \mid x_2)$ are at least $m - \alpha(n) - 7 \log n$. We show this relation for $K(x_j \mid x_2)$ (the proof for $K(x_j \mid x_1)$ is similar). Suppose that $K(x_j \mid x_2) < m - \alpha(n) - 7 \log n = m - t$. Let $t_1 = K(x_1)$. Note that $t_1 \geq s(n)$. Let $B = \{u \in \{0, 1\}^n \mid K(u) \leq t_1\}$. Note that $2^{t_1+1} > |B| \geq 2^{2s(n)/3} = S$. (B has size $\geq 2^{2s(n)/3}$ because it contains the set $0^{s(n)/3} \{0, 1\}^{2s(n)/3}$.) We say that a column $u \in [N]$ is *bad for color* $a \in [M]$ and B if the number of occurrences of a in the $B \times \{u\}$ subrectangle of T is greater than $(2/M) \cdot |B|$ and we say that u is *bad for* B if it is bad for some color a and B . For every $a \in [M]$, the number of u 's that are bad for a and B is $< S$ (since T is (S, n^k) -strongly balanced and we can take into account the first balancing property of such tables). Therefore,

the number of u 's that are bad for B is $< M \cdot S$. Given t_1 and a description of the table T , one can enumerate the set of u 's that are bad for B . This implies that any u that is bad for B can be described by its rank in this enumeration and the information needed to perform the enumeration. Therefore, if u is bad for B ,

$$\begin{aligned} K(u) &\leq \log(M \cdot S) + 2(\log t_1 + \log n) + O(1) \\ &\leq m + 2s(n)/3 + 4 \log n + O(1) \\ &< s(n), \end{aligned}$$

provided n is large enough. Since $K(x_2) \geq s(n)$, it follows that x_2 is good for B .

Let $A = \{w \in [M] \mid K(w \mid x_2) < m - t\}$. We have $|A| < 2^{m-t}$ and, by our assumption, $x_j \in A$. Let G be the subset of B of positions in the strip $B \times \{x_2\}$ of T having a color from A (formally, $G = \text{proj}_1(T^{-1}(A) \cap (B \times \{x_2\}))$). Note that x_1 is in G . Each color a occurs in the strip $B \times \{x_2\}$ at most $(2/M) \cdot |B|$ (because x_2 is good for B). Therefore the size of G is bounded by

$$|A| \cdot (2/M) \cdot |B| < 2^{m-t} \cdot (2/M) \cdot 2^{t_1+1} \leq 2^{t_1-t+2}.$$

Given $x_2, t_1, m - t$ and a description of the table T , one can enumerate the set G . Therefore, x_1 can be described by its rank in this enumeration and by the information needed to perform the enumeration. It follows that

$$\begin{aligned} K(x_1 \mid x_2) &\leq t_1 - t + 2 + 2(\log t_1 + \log(m - t) + \log n) + O(1) \\ &\leq t_1 - t + 6 \log n + O(1) \\ &= t_1 - \alpha(n) - \log n + O(1) \\ &= K(x_1) - \alpha(n) - \log n + O(1), \end{aligned}$$

which contradicts that x_1 and x_2 have dependency at most $\alpha(n)$.

We next prove Claim 2.

We fix two elements $i \neq j$ in $\{3, \dots, k + 2\}$ and analyze $K(x_i \mid x_j)$.

Let $t_1 = K(x_1)$ and $t_2 = K(x_2)$. From hypothesis, $t_1 \geq s(n)$ and $t_2 \geq s(n)$. We define $B_1 = \{u \in \{0, 1\}^n \mid K(u) \leq t_1\}$ and $B_2 = \{u \in \{0, 1\}^n \mid K(u) \leq t_2\}$. We have $S \leq |B_1| < 2^{t_1+1}$ and $S \leq |B_2| < 2^{t_2+1}$. (B_1 and B_2 have size larger than $S = 2^{2s(n)/3}$, because they contain the set $0^{s(n)/3}\{0, 1\}^{2s(n)/3}$.)

Let $T_{i,j}^{-1}(x_i, x_j)$ denote the set of pairs $(u, v) \in [N] \times [N]$ such that $T(u + i, v) = x_i$ and $T(u + j, v) = x_j$.

Note that $(x_1, x_2) \in T_{i,j}^{-1}(x_i, x_j) \cap (B_1 \times B_2)$. Since the table T is strongly balanced,

$$|T_{i,j}^{-1}(x_i, x_j) \cap (B_1 \times B_2)| \leq \frac{2}{2^{-2m}} 2^{t_1+t_2+2} = 2^{t_1+t_2-2m+3}.$$

Note that $T_{i,j}^{-1}(x_i, x_j) \cap (B_1 \times B_2)$ can be effectively enumerated given x_i, x_j, i, j , and the table T . Thus $x_1 x_2$ can be described from $x_i x_j$, the rank of (x_1, x_2) in the above enumeration, i, j , and the table T . This implies that

$$\begin{aligned} K(x_1 x_2) &\leq t_1 + t_2 - 2m + 3 + K(x_i x_j) + 2k \log n + 2(\log k + \log \log n) + O(\log n) \\ &\leq t_1 + t_2 - 2m + K(x_i x_j) + (2k + O(1)) \log n. \end{aligned}$$

On the other hand, $K(x_1x_2) \geq K(x_1) + K(x_2 \mid x_1) - O(\log n)$ and $K(x_2 \mid x_1) \geq K(x_2) - \alpha(n)$. Therefore,

$$\begin{aligned} K(x_1x_2) &\geq K(x_1) + K(x_2) - (\alpha(n) + O(\log n)) \\ &= t_1 + t_2 - (\alpha(n) + O(\log n)). \end{aligned}$$

Combining the last two inequalities, we get that

$$t_1 + t_2 - (\alpha(n) + O(\log n)) \leq t_1 + t_2 - 2m + K(x_ix_j) + (2k + O(1)) \log n$$

which implies that

$$K(x_ix_j) \geq 2m - \alpha(n) - (2k + O(1)) \log n.$$

Therefore

$$\begin{aligned} K(x_j \mid x_i) &\geq K(x_ix_j) - K(x_i) - O(\log n) \\ &\geq (2m - \alpha(n) - (2k + O(1)) \log n) - (m + O(1)) - O(\log n) \\ &= m - \alpha(n) - (2k + O(1)) \log n. \end{aligned}$$

It follows that

$$\begin{aligned} K(x_j) - K(x_j \mid x_i) &\leq (m + O(1)) - (m - \alpha(n) - (2k + O(1)) \log n) - O(\log n) \\ &\leq \alpha(n) + (2k + O(1)) \log n. \end{aligned}$$

Thus, x_j and x_i are at most $\alpha(n) + (2k + O(1)) \log n$ -dependent. ■

3 Polynomial-time generation of one random string

In this section we prove Theorem 3. The formal statement is as follows.

Theorem 3. *For every $\delta > 0$ and for every function $\alpha(n)$, there exists a constant c and a polynomial-time computable function $f : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}^m$, where $m = \Omega(\delta n)$, with the following property. If n is sufficiently large and x and y are two strings of length n satisfying*

- (i) $K(x) \geq \delta n$
 - (ii) $K(y) \geq \delta n$
 - (iii) x and y are at most $\alpha(n)$ - dependent,
- then

$$K(f(x, y)) \geq m - (\alpha(n) + O((\log n)^c)).$$

Proof. Let $\epsilon = 1/(8n^{10} \cdot \alpha(n))$ and let c be the constant and $Ext : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}^m$ be the function given by Theorem 1 for parameters $(\delta/2)$ and ϵ . Let $t = \alpha(n) + 10 \log n + ((\delta/2) \log 1/\epsilon)^c + 3 = \alpha(n) + O((\log n)^c)$.

The function f on input x and y returns $z = Ext(x, y)$. We show that $K(z) \geq m - t$.

Suppose $K(z) < m - t$.

Let $t_1 = K(x)$, $t_2 = K(y)$, $B_1 = \{u \in \{0, 1\}^n \mid K(u) \leq t_1\}$, $B_2 = \{u \in \{0, 1\}^n \mid K(u) \leq t_2\}$. From hypothesis, $t_1 \geq \delta n$ and $t_2 \geq \delta n$.

Note also that $2^{\delta n/2} \leq |B_1| \leq 2^{t_1+1}$ and $2^{\delta n/2} \leq |B_2| \leq 2^{t_2+1}$. (The sets B_1 and B_2 have size $\geq 2^{\delta n/2}$ because they contain $0^{n-\delta n/2}\{0,1\}^{\delta n/2}$.)

Let

$$A = \{v \in \{0,1\}^m \mid K(v) < m - t\}$$

We focus on the table defined by the function $Ext : [N] \times [N] \rightarrow [M]$, where, as usual, we have identified $\{0,1\}^n$ with $[N]$ and $\{0,1\}^m$ with $[M]$.

Let G be the subset of $B_1 \times B_2$ of cells in the rectangle $B_1 \times B_2$ that are A -colored.

Since $Ext(x, y) = z \in A$, $x \in B_1$ and $y \in B_2$, the cell (x, y) belongs to the rectangle $B_1 \times B_2$ and is A -colored. In other words, $x \in G$.

Taking into account Lemma 2, we can bound the size of G by

$$\begin{aligned} & \left(\frac{|A|}{M} \cdot 2^{((\delta/2) \log 1/\epsilon)^c} + \epsilon \right) |B_1 \times B_2| \\ & \leq 2^{t_1+t_2+2} \left(\frac{2^{m-t}}{2^m} \cdot 2^{((\delta/2) \log 1/\epsilon)^c} + 2^{-\log 1/\epsilon} \right) \\ & = 2^{t_1+t_2-t+((\delta/2) \log 1/\epsilon)^c+2} + 2^{t_1+t_2-\log 1/\epsilon+2} \\ & \leq 2^{t_1+t_2-(\alpha(n)+10 \log n)}. \end{aligned}$$

The last inequality follows from the choice of ϵ and t .

The set G can be enumerated if we are given t_1 , t_2 , δ and n (from which we can derive t and the table Ext), and every element in G can be described by its rank in the enumeration and by the information needed to perform the enumeration.

Since $x \in G$, it follows that

$$\begin{aligned} K(xy) & \leq t_1 + t_2 - \alpha(n) - 10 \log n + 2(\log t_1 + \log t_2 + \log n) + O(1) \\ & < t_1 + t_2 - \alpha(n) - 4 \log n \\ & = K(x) + K(y) - \alpha(n) - 4 \log n. \end{aligned}$$

We have used the fact that $t_1 \leq n + O(1)$ and $t_2 \leq n + O(2)$. By the Symmetry of Information Theorem,

$$K(xy) \geq K(y) + K(x \mid y) - 2 \log n - O(1).$$

Combining the last two inequalities, we get

$$K(x) - K(x \mid y) > \alpha(n) + \log n,$$

which contradicts the fact that x and y are at most $\alpha(n)$ -dependent. ■

References

- Bou05. J. Bourgain. More on the sum-product phenomenon in prime fields and its applications. *International Journal of Number Theory*, 1:1–32, 2005.
- CZ08. C. Calude and M. Zimand. Algorithmically independent sequences. In *Developments in Language Theory*, volume 5257 of *Lecture Notes in Computer Science*, pages 183–195. Springer, 2008.

- FHP⁺06. L. Fortnow, J. Hitchcock, A. Pavan, N.V. Vinodchandran, and F. Wang. Extracting Kolmogorov complexity with applications to dimension zero-one laws. In *Proceedings of the 33rd International Colloquium on Automata, Languages, and Programming*, pages 335–345, Berlin, 2006. Springer-Verlag *Lecture Notes in Computer Science* #4051.
- Rao08. Anup Rao. A 2-source almost-extractor for linear entropy. In Ashish Goel, Klaus Jansen, José D. P. Rolim, and Ronitt Rubinfeld, editors, *APPROX-RANDOM*, volume 5171 of *Lecture Notes in Computer Science*, pages 549–556. Springer, 2008.
- Raz05. Ran Raz. Extractors with weak random seeds. In Harold N. Gabow and Ronald Fagin, editors, *STOC*, pages 11–20. ACM, 2005.
- VV02. Nikolai K. Vereshchagin and Michael V. Vyugin. Independent minimum length programs to translate between given strings. *Theor. Comput. Sci.*, 271(1-2):131–143, 2002.
- Zim08. Marius Zimand. Two sources are better than one for increasing the Kolmogorov complexity of infinite sequences. In Edward A. Hirsch, Alexander A. Razborov, Alexei L. Semenov, and Anatol Slissenko, editors, *CSR*, volume 5010 of *Lecture Notes in Computer Science*, pages 326–338. Springer, 2008.
- Zim09. M. Zimand. Extracting the Kolmogorov complexity of strings and sequences from sources with limited independence. In *Proceedings 26th STACS, Freiburg, Germany*, February 26–29 2009.
- ZL70. A. Zvonkin and L. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 25(6):83–124, 1970.